

IASSIST

QUARTERLY

Volume 10 Number 4

Winter 1986

FEATURES

- 3** **Directions of Major Archives**
by Bjorn Henrichsen
- 7** **Interactive Access to Survey Databases**
by Mark Katz and Beverly Rowe

DEPARTMENTS

- 19** **News & Notes**
- 21** **Current Journal Contents**

IASSIST

QUARTERLY

Volume 10 Number 4

Winter 1986

FEATURES

- 3** **Directions of Major Archives**
 by Bjorn Henrichsen
- 7** **Interactive Access to Survey Databases**
 by Mark Katz and Beverly Rowe

DEPARTMENTS

- 19** **News & Notes**
- 21** **Current Journal Contents**

Editorial Information

The IASSIST QUARTERLY represents an international cooperative effort on the part of individuals managing, operating, or using machine readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

Information for Authors

The QUARTERLY is published four times per year. Articles and other information should be typewritten and double-spaced. Each page of the manuscript should be numbered. The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press *Manual of Style* or Kate L. Turabian's *Manual for Writers*. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. Bibliographic references for numeric social science data files: suggested guidelines. *Journal of the American Society for Information Science* 30(2):77-82, March 1979. If the contribution is an announcement of a conference, training session, or the like, the text should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event. Book notices and reviews should not exceed two double-spaced pages. Deadlines for submitting articles are six weeks before publication. Manuscripts should be sent in duplicate to the Editor:

Walter Piovesan
Research Data Library
W.A.C. Bennett Library
Simon Fraser University
Burnaby, B.C., V5A 1S6 CANADA
(01)604/291-4349 E-Mail: Piovesan@SFU.MAILNET

Book reviews should be submitted in duplicate to the Book Review Editor:

Kathleen M. Heim
School of Library and Information Science
Louisiana State University
Coates Hall, Room 267
Baton Rouge, Louisiana 70803 USA
(01)504/388-3158

Key Title: Newsletter - International Association for Social
Science Information - Service and Technology

ISSN - United States: 0739-1137 Copyright © 1985 by IASSIST. All rights reserved.

Digitized by the Internet Archive
in 2010 with funding from
University of North Carolina at Chapel Hill

<http://www.archive.org/details/iassistquarterly104inte>

Directions of Major Archives

by Bjørn Henriksen¹

Let me start by giving a broad overview of the organizational structure and main services provided by the Norwegian Social Science Data Services (NSD). Based on this description I will then close by saying a few words about new services to be developed in the coming years.

NSD was formally established in 1971 as an organ of the Norwegian Research Council for Science and the Humanities (NAVF).

NSD differs from most similar organizations in five ways:

- it is a federally structured facility with offices at all four universities in Norway and in the regional colleges at smaller centers across Norway. Its headquarters are at the University of Bergen;
- it has built up a wide variety of data resources in all fields of the social sciences: not only data from surveys, but also a large data bank for communes and census tracts, an archive of information about organizations, and a series of files on the

recruitment and careers of various elite groups;

- it acts as the Census Bureau's distribution agency to the academic community;
- it has set up a special service responsible for contacts between the research community and the Governmental Data Inspectorate; and
- it has established a national service for information on current research in the social sciences.

In comparison with most other data facilities established in Europe and in the U.S. in the last two decades, NSD is probably the one giving the highest priority to book-keeping and "process produced" data. It is deliberately multisectoral and sees its primary task as to link up and to systematize data of different types; in contrast to the typical survey archive, it is not just a repository of separately documented data sets. It is even correct to say that it is only in the last few years that NSD has been active in archiving data from various research projects.

Among the larger data holdings of the NSD are:

The Commune Data Base 1769 - 1986

This data base contains statistics on all local administrative units in Norway since 1769, and is linked up with a computer cartography facility. This is the most widely used facility in the NSD and is constantly expanded and improved. A great deal of energy has been invested in developing effective solutions to the problems posed by changes in boundaries and in the number of units. The base includes detailed documentation of all such changes that have taken place.

¹Prepared for Opening Plenary Session, IASSIST Conference, Marina Del Rey, May 1986

Coordinate matrices for all commune boundaries have been established, and boundary segments

are time coded to allow the production of maps for the units existing at any particular time period since 1769.

As of 1986 the Commune Data Base includes about 29,000 variables for each commune.

Census Tract data base 1950 - 1980

To allow analyses at a lower level of aggregation, NSD has also organized a system of data for the lowest level of official enumeration: the census tract. This data base includes the censuses of 1950, 1960, 1970 and 1980.

Census Data Bank 1960-1970-1980

10% of the population are followed through three censuses. The data base includes approximately 483,000 individuals.

Nordic Regional Data Base

The Social Science Research Councils of Denmark, Finland, Norway and Sweden have funded this data base.

Data are gathered and organised in systematic time-series for all five Nordic countries, including Iceland. The regional units of the data are counties:

amt for Denmark
län for Sweden and Finland
fylker for Norway
syslu for Iceland

Time series are created for all units from 1850 to 1980 for population census data, and the period 1945 to 1980 for other groups of data.

The data base system is composed of four elements:

- A set of data for the post-war period, consisting of 4-500 variables for each unit,

most of them organized in five-year time series:

- A longitudinal set of data based on population censuses 1850-1970/80 consisting of 100-150 variables organized in ten-year time series;
- A data set on population movements 1945-1980 consisting of an annual time series for each unit;
- A set of coordinate matrices for the boundaries of the units. This includes time-specific segments wherever there have been changes in the boundaries of units in the period from 1850 to 1970/80.

Criminal Justice Data

NSD also has an archive of Norwegian criminal justice data from 1860 to 1975.

Gallup Data

This collection is based on data from Norsk Gallup Institutt and Norsk Opinionsinstitutt. It contains their monthly surveys from 1964 to the present.

Election Studies

NSD has taken over the surveys conducted by the Norwegian Election Project. Data from the following national surveys are available from NSD: 1957, 1965, 1969, 1977, 1981.

Surveys from the Central Bureau of Statistics

Some of the most thorough surveys in Norway have been carried out by the Central Bureau of Statistics from 1967 to the present. The data from these surveys are at the disposal of academic users in Norway via NSD.

Members of Parliament

A data bank has been established containing information about all Members of Parliament and the Government. It covers the period from 1814 and includes information on father's occupation, education, early career, positions in legislative committees, etc.

Members of Official Committees

This collection includes information on all committees appointed by the various Ministries, as well as the members of such committees. It covers 1936, 1951, 1966, and every year from 1980 on.

Voluntary Associations

The file includes data on the 1300 largest voluntary associations in Norway. Data are available for the following years: 1964, 1967, 1970, 1976 and 1983.

Teaching Packages

NSD has given priority to the establishment of a set of teaching packages for both the universities and the regional colleges. In 1985, we also launched a program to establish working tools for the Norwegian high schools. The program has been accepted and is financed by the Norwegian Ministry of Education. Our first products under this program are now in use in the Norwegian schools.

Of the new NSD services established in the last five years, I will mention two:

Secretariat for Data Protection Affairs

The Norwegian Personal Data Registers Act (Lov om person-registre m.m.) came into force in 1980. In response to proposals from the Social Sciences, the Research Council in 1980 established the Secretariat for Data Protection Affairs as a part of NSD. The Secretariat was accepted as a broker between the research community (including medicine, the humanities,

etc.) and the Data Inspectorate, and was mandated to provide regular reports to the Data Inspectorate on all projects funded through the Research Council for which concession was required in accordance with the provisions of the Act. Since then, the Secretariat has been given the same mandate for all research carried out at the universities with grants from other sources than the Research Council.

Through agreements between the Data Inspectorate, the Research Council and the universities, NSD has also been given the responsibility of archiving data, provided there is reason to assume their usefulness in future research.

Information Service for On-Going Research

In 1984, the Research Council established an information service for Norwegian research, the aim of which is to improve awareness of current research; it is provisionally established for a period of five years.

The Information Service, in addition to general management, consists of one branch responsible for research in the Humanities, and one branch responsible for research in the Social Sciences. The Social Science branch is located at NSD.

The Service is active in all fields in which the Research Council is engaged, i.e. medical science, the humanities, social science and research for social planning.

The information is available in a data base for convenient access from users' own terminals. In addition there are printed catalogs for specific research areas.

A small country with 4.5 million inhabitants and only four universities must coordinate national activities. For NSD it has meant that not only the Research Council but also the universities

and to a certain extent the regional colleges have chosen to concentrate their means for a social science infrastructure at NSD. Today, different parts of the Research Council cover about 75% of our expenses, while the universities and research projects cover the rest. We have today a professional staff of 15 and 4 clerical staff members. Including assistants etc., we estimate that about 28 full-year equivalents will be utilized in 1986.

Given our special relationships to the Research Council, the universities, the Census Bureau, and the Data Inspectorate, we have today a monopoly in the areas in which we are active.

We are striving to fulfill our responsibilities, to serve our users in the best possible way, and to make our services easily accessible to the scientific community. Until now, our data services have been made operative through local offices at each of the universities, located in the University computer centers. The network among the universities has not been seen as an alternative to direct service with our own staff present at the local university. Initiatives have, however, now been taken to make the network function better, and we expect that, within two years, some of our data holdings will be held in Bergen only, to be requested via the university network for local users at other institutions.

The services we provide today cover a broad range: from a data base with information on all social science projects and publications based on them, to our own data banks and data from all projects financed by the Research Council, to projects financed by other institutions, such as the universities and some of the ministries.

Given that researchers must deposit their data with NSD, they are informed of standards for documentation and data, which means a standardization among widely separated scholars. As the data holdings grow, there is an increased need for researchers to be kept informed of the data holdings and services. We act not only as

a distributor of data, but also as a broker of social science information.

We think that we have played an important role in giving researchers easy access to information and in introducing new technology. Through our work we have prevented duplication of data work, we have made data available free to various users, and we have stimulated cumulative research by making data from earlier projects available to new ones. Although our data have been mainly used in the social sciences, users in other fields such as history, medicine etc. are increasingly using our services. Our greatest growth potential within the research community lies in serving these new groups. We are on our way from being a social science service to being a more general service for a broader group of users.

In the past, we have concentrated our efforts on providing service to the research community. During the last few years we have also started to serve local and federal agencies. These are now in the same position as the social science community five years ago, and they now want access to the services established for researchers. New recruits to governmental agencies often find that they, in their new position, do not have the easy access to data they had had as students. As students they were introduced to our services and now, they are still in need of access. We are discussing ways of serving both researchers and bureaucrats, and believe that we will agree on a model covering both needs. Presently we are also negotiating with the Norwegian Parliament to make our services available to both members of Parliament and their staff. By pooling resources from these different sources, all parties will have access to a much broader range of services. Our main efforts in the coming few years will be devoted to planning a shared information system for planners and researchers, and hopefully we can, within a few years, present a system serving a broader community than today.□

Interactive Access to Survey Databases

by Mark Katz and Beverley Rowe
QUANTIME Limited, London, UK

Scenario

As a data-archivist, you have within your computer library tens, perhaps hundreds, of surveys. Many of these are heavily used and you can afford to have them on-line on your small mini-computer. Lack of funds has prevented you from installing CD-ROM or laser disks and you have only a few programmers.

The phone rings: someone looking into the effects of radio-active fallout wants a quick statistic from one of your on-line surveys - one you are not too familiar with. It requires a scan over three years of some 60,000

incompatible questionnaires to select a sub-group of all people who studied physics at university and may have contracted cancer.

Five years ago you would have sent them a tape of data, possibly also some SPSS control commands, and told them to do it themselves, a task that would have taken them days or even weeks to complete. Last year you could have sent them a floppy disk but still a long and daunting task would lie ahead of them.

Now you log into your computer, access the survey and, even though you do not have the questionnaires

handy, find the name of all variables which look at 'physics' and 'cancer'. Within one minute, you dictate the relevant statistics over the phone.

The caller is most impressed with the speed and is quite interested in the results - "How can I find out more?" he asks. "Either dial in to the database or, if you have a PC, I can provide you with the total database on just a few floppies. It's the same interactive system, designed for very fast access by researchers with no previous training" you reply.

An ideal view of the future? No, Quantime offers this type of service today.

The market research industry conducts thousands of surveys each year. Traditionally, the results of surveys are produced as hard copy computer tabulations, but this is changing, and more surveys are ending up as on-line databases.

Quantime is in the forefront of such developments with a user-friendly system called QUANVERT which offers fast access to such databases, some of which exceed two million cases/respondents.

This paper reviews our experiences with offering remote access to large survey databases and working with the data archive at the University of Essex to promote the use of on-line access to the General Household and other surveys.

QUANVERT

Background

Quanvert is the interactive member of a family of software tools for processing survey data,

offering non-technical users direct access to data with an interface especially designed for them. It permits very fast exploratory access to data by taking advantage of inverted or transposed file structures.

Quanvert is five years old, written in C and currently runs on DEC-Vac (Unix and VMS), Prime, large IBM mainframes (MVS, VM/CMS), many Unix-based micros and, more recently, the IBM PC/AT.

Quanvert handles numeric, categoric, multi-coded and textual variables with all normal boolean and arithmetic functions. Users may cross-tabulate or interrogate data as well as create new variables, all interactively. Increasingly, users are linking to Quanvert through PC's to download data and use spreadsheets and graphics. An associated package, QUANTUM, sets up the data description.

Some facilities

Quanvert reads transposed files to produce data at either the *aggregate* level (as cross-tabulations) or *disaggregate* level (specific values/responses for selected records).

It interacts with the user to determine the variables to be selected and the types of reports to be produced. The variables (or axes) corresponding to, or derived from, the original fields/questions may be manipulated, tabulated, displayed or used in statistical analysis.

Sub-sets of data

Subsets of the data are extracted by *filter* commands that use logical or arithmetic combinations of existing variables.

The specification of these *filters* processes code text rather than data values. For instance, selecting only women uses the variable *sex* and subset *female* rather than looking at bytes 4-6,

value 2, etc. For frequently used selections a named filter may be created, but Quanvert does not set aside physical subsets.

Types of data

- * Categorical (sex, region, etc.)
- * Multi-coded (makes of car owned, etc.)
- * Numeric (salary, date, etc.)
- * Alphabetic (names/addresses, verbatim responses and text)
- * Hierarchical (master/trailer)

This last facility allows for analysis at different levels of data for accumulation across levels.

Types of report

- * Simple cross-tabulations (up to six dimensions)
- * Filtered tabulations using logical combinations of variables
- * Means or proportions and table division
- * Grossed-up tables (multiple weights if needed)
- * Listings of raw data

Operations to look after the database

- * Create new variables
- * Delete/rename variables
- * Create special filters

- * Combine similar data for months/areas, etc.
- * Join data from different surveys
- * Manipulate variables across levels in a hierarchy
- * Print a code book, including KWIC index of the database text

Help commands

- * Lists of commands and variables
- * Detailed explanations
- * Marginals (summary statistics) for each variable
- * Text search for keywords in the code book

Other features

- * Statistical analyses
- * Combinatorial analysis
- * Sorted/accumulated lists
- * Production of sticky address labels
- * Data downloading for a micro
- * Files for Symphony/Lotus, etc.

Perhaps the most powerful facility is that separate surveys can be stored as individual data sets and then 'joined' together. Thus, data for different years may be aggregated to compare results over time. Quanvert automatically introduces a new variable (*years* or whatever the appropriate unit) which may be used as a

breakdown. Quanvert looks after minor changes between questionnaires from different years. Providing the code text and options remain constant, Quanvert transparently combines data even though the position on the questionnaire has altered.

Data reading time depends on the degree of filtering. Unfiltered requests are processed at speeds of 500-1000 cases per second, irrespective of the size of database or number of variables. On the Vax 750, it is not uncommon to reach speeds of up to 30,000 cases per second on heavily filtered tables. Even on the Compaq (IBM/AT compatible PC) Quanvert processes up to 15,000 cases per second. Speeds of up to 200,000 respondents per second have been recorded on an IBM mainframe.

Defining the Data

Introduction to QUANTUM

Quanvert is closely linked to the package Quantum. This batch-oriented program edits, recodes, and tabulates survey data. The user sets up a specification file which describes the data together with the recoding and analyses required. This file is compiled by Quantum and run on the original raw data file, a multi-stage process involving data reading, data accumulation and report printing.

A typical Quantum specification file has two sections:

The *EDIT* section uses a special language that combines many features of Fortran with special facilities for handling survey and structured data. It includes powerful data checking commands and an online data correction facility.

The *TABULATION* section contains non-procedural statements that:

- a. Define the axes. For instance, the following statement specifies a variable *Sex* which may be found on position 6 of the data file, where 1 denotes *Male* and 2 denotes *Female*

```
l sex
col6;hd=Sex of Respondent;Base=Total
Sample;Male;Female
```

- b. Define the tabulations, specified as a series of *TAB* statements. These use the predefined axes as rows, columns and filters. This section offers a large selection of options for format control for mathematical computations and percentage calculation, as well as detailed layout of headings, row/column text, figures and labelling.

Setting up the data description for Quanvert

Quantum is used for this. The user specifies the variables plus associated headings, text and location using the tabulation section. Any recording or derivation of new variables would be included in the edit section.

Quantime has developed a semi-automated SPSS-Quantum conversion package.

Preparing the transposed file

The *flip* program is now invoked to read the Quantum specification, extract and recode data from the original data file and prepare the transposed file. Since the transposed file contains the original data *and* the data description, it is not necessary to retain the original data files. The time taken to invert the General Household Survey (12,000 households,

23,000 people and some 120 variables) was only two hours on the Compaq. The Appendix contains details of this transposed file.

Extending the Data

Even though Quanvert works with transposed files, it is possible to add new cases or variables. It is rarely necessary to go back to Quantum to create variables.

Shorthand methods are provided to copy a variable with the addition of an existing filter or set of filters. The new variable becomes part of the database.

More generally, the user sets up in a separate directory a mini database containing the new records and this is added to the database. It is not necessary to reprocess the entire database. Secondary databases are simply appended to the main database, i.e. each file in turn is appended to the relevant variable-file.

However, in many situations it is better to keep additional sets of records separate. For instance, data may arrive in monthly batches or from different areas. In this case all the secondary databases are *joined* together in a *MULTI-FLIP* structure, such that there is one master directory and multiple subdirectories. Quanvert creates a new variable which contains as elements each of the subdirectories, e.g., month or country. This allows the user to tabulate any variable by, for example, month or select any number of sub-directories for an analysis.

Multi-flip looks after changes to the variables between batches of data. If the number of elements and the code text are unchanged (even if they come from different parts of the questionnaire), those variables are assumed to be accessible to all sub-directories.

If a new *variable* needs to be added to the database (or an existing one replaced), it is not necessary to set the database up again. The user may create a new variable directly within Quanvert, using logical combinations of existing variables. Alternately, if new data have been provided or additional external variables are required, these may be prepared separately and *merged* into the main database. This will add or replace those with identical names.

Post-processors

Quanvert has facilities to select variables from specified respondents and to write this out to a file. This file may then be downloaded to another system for statistical or graphic operations. An option provided will convert the values of variables into numeric fields, rather than the text of the value (e.g. value 1 for *male* and 2 for *female*) and thus simplify the interface to statistical systems. This option also provides the SPSS variable and value labels. A useful facility on Unix-based systems is to pipe this output to user defined post-processors directly rather than to an external file.

Post-processors are provided to reformat cross-tabulations into a form acceptable to other packages. This uses the SYLK file-format or shortened character format for input to Symphony/Lotus with the *FILE-IMPORT* option.

To summarise, then, Quanvert offers very fast analysis of survey data. It combines simplicity of use with a wide range of facilities. The interface to other packages, its linkage facilities to download to micro-computers and its availability on a broad range of computers, from large IBM, most major mini-computers, to the IBM-PC, makes it a leading package for the analysis of large survey databases.

What is the GHS?

The General Household Survey (GHS) is one of many surveys conducted each year by OPCS (Office of Population Census and Surveys) in London, a government department. It is considered a cornerstone of social research in the UK.

The GHS is carried out each year with some 12,000 households/23,000 people as a hierarchical data set. It is normally available within 6 – 9 months of the end of fieldwork.

The survey covers a broad range of topics: health, education, car ownership, use of energy, employment, income, family size, age, and so on. There are some 800 basic variables.

OPCS carry out the data collection, cleaning and preliminary analysis. They have switched to Sir for data management but still use a fairly old system on their ICL computer for the main reporting.

A Monitor appears each year as the first indicator of social change but the OPCS is unable to provide much or fast response for further reporting. They use the Data Archive at the University of Essex as a distribution point and invite bona fide researchers to conduct any further work themselves on their own computers, using the raw data.

The ESRC Data Archive

The Data Archive at the University of Essex is funded by the ESRC (Economic and Social Research Council) and is one of the largest collections of machine-readable survey data files in Europe. They have thousands of surveys, many with an associated SPSS control files. They publish a quarterly newsletter and hold an important position in the social survey world. Many surveys from the private sector are held

by the Archive, and it is a condition of all ESRC grants that resulting data are deposited there.

The GHS Experiment

By mid-1984, Quantime had considerable experience with remote databases and a well-established UK-based service for the private sector. Quantime felt that this concept needed to be introduced to the public and academic sectors and initiated discussions with the Archive to take an important and well used dataset for implementation as part of the Quanline service. After lengthy discussion and approval from OPCS, it was decided to use three years of GHS data, and work began in mid-1985.

Quantime set up this important public database and made it available at no charge to bona fide researchers in the academic sector through the Quanline time-sharing service. Agreement was reached whereby Quanline joined with Essex in low-level marketing to the academic and public sectors. It was hoped that the experience would help the Data Archive in any plans to make datasets available interactively, rather than by mailing tapes or floppy disks.

After unsuccessful attempts to obtain external funding, Quantime allocated over 50,000 dollars to the project from internal funds. This included recruiting a consultant to develop the database and market the concept as well as an allowance of computing resources to store the data, set up the database and provide free access time to users.

Raw data and a list of variables were supplied in July 1985. At the time, we were unable to obtain an SPSS file for this dataset and it had to be set-up 'manually' in Quantime, a daunting task. The variables were prepared and the Quanvert database was available in September 1985.

A subset of the data for 1980, 1981 and 1982 was put up, including most of the household level data and important parts of the individual level data; in all, some 120 variables of a total of 600. However, the choice of variables turned out to be a poor one and insufficient key topics were available. The database is available on the Vax under Unix as a time-sharing service through Quanline. The user has access to all years and may produce comparative reports between years. The 1980 data are also available on the IBM PC/AT (and Compaq). On all machines, the average time to scan one year's data is under 10 seconds for either household-level or individual-level reports.

GHS Data On-Line

Progress in marketing has been steady. Contact, sometimes to considerable depth, has been made with over fifty academic, public sector and quasi-public consultancy organisations. Marketing has focused on mailshots, telephone calls, direct mail and press releases.

This has led to very positive interest in the public sector and given an interesting insight into the (largely unsatisfied) demand for GHS data, into the research and thinking of users attempting to obtain statistical information from large survey databases, and the constraints under which they operate.

University and polytechnic users are being offered Quanvert for GHS at no charge. There are now eleven committed users at six sites.

The main problems with promoting the current version of GHS have been:

- the non-coverage of important areas of interest
- rather old data.

Whether interested in research or reference, the user must be able to find everything that was collected. The lack of income variables in the earlier releases of the service was particularly disabling, but other areas have proven important to particular users.

We can expect broadly two uses of the GHS or other large survey databases: active research and casual reference. We would expect academics to fall in the first category, non-academic researchers (public or private sector) in the second. Because of the computing and staff resources required to obtain information quickly from archived survey data, most people terminate their research prematurely, or turn to other sources (often at great cost) to find information that is duplicated in GHS.

Discussions are now taking place to open up GHS data to the private sector and to charge for this service.

Other datasets were selected to supplement the GHS, namely the WFS (World Fertility Survey) Fiji survey of 4,900 respondents and 300 variables, and NCDS (National Child Development Survey) of some 18,000 children and 350 variables. We hope to have the three-year British Social Attitudes Survey on-line before the end of 1986. These surveys run alongside other private datasets resident on the Quanline computers and include: British

Telecomm's Telecare project (data from 3 million respondents); Manpower Services Commission (400,000 Youth Training Scheme trainees); British Gas's NDES project (55,000 establishments).

All of these are accessed by regional marketing and research staff at hundreds of offices around the UK.

Although we have dedicated this section of the paper to our work with the Data Archive, it represents less than 5% of the work of Quanline

UK, measured in terms of computing resources and usage by researchers.

A Summary of our experiences

Some Observations

We are able to assess the impact of interactive survey analysis based on our experience of some five years of Quanvert, many hundreds of users and some 2-3,000 connect hours each month on databases ranging in size from a few hundred to a few million cases.

a. Users cannot grasp the concept

Since most researchers are unable to obtain really fast or simple access to large databases, ad-hoc or reference interrogation has been largely overlooked. It cannot be understood without a demonstration or trial evaluation and an element of retraining. It is so alien to most people that there is a barrier to its introduction. They say: *What extra benefit is there to me if the results come back in two minutes instead of two hours?*

b. Software development is misdirected

A lot of human resources are spent on developing easy analysis systems, but not enough on good data organisation or easy data access. User-friendliness only comes with many users and long sessions by people other than the primary user or programmer. The use of 'laser disks' demands a new type of storage mechanism. If they are to be used speedily and effectively, we cannot simply replace the old floppy or Winchester disk but use the same old software.

Many people are using tailor-made programs and re-inventing wheels in software development. There are too many government departments using (and even re-writing) Cobol programs for survey analysis.

Archived data require a *read-only* strategy. Conventional DBMS programs place too much emphasis on updating rather than (fast) reporting. They are also greedy for computing and storage requirements.

c. The need for Statistics is exaggerated

The importance of statistical reporting is over-emphasised; in fact it represents less than 20% of access. The 80% can be achieved by (complex) cross tabulations. Despite this, the availability of good statistics seems to be a more important criterion than speed, flexibility or user-friendliness. In practice a 'database server' is required as a front end to the statistical software.

What Changes are Necessary?

There are now some 2,500 bibliographic and textual databases available. Users spend millions of dollars each year; an entire industry has been built up, with conferences, training, newsletters, books and software investment. But this information is mostly textual and is difficult to manipulate arithmetically. It is also at a very high level of aggregation.

As the provision of fast, interactive tabulations from large databases gains momentum, a number of key issues are evolving:

- a. Data consistency is important. Data under intensive scrutiny will lose credibility if badly formed.
- b. Help information is needed, down to the

variable level.

- c. A support team must be available to deal with queries on computing, telecommunication, and data problems.
- d. A database of databases is required to indicate the best source for information.
- e. There must be a common format for a data description language so that users can provide translators to/from a common language. Unfortunately this probably has to be SPSS, but a more comprehensive dictionary approach would be better.
- f. Users must stop developing their own special analytical tools and rely on those already in use. The public sector must be prepared to go to the private sector and to scour the world for the best system.
- g. The main software developers must turn to inverted files as a basis for fast access.
- h. Specialised software on micros (spreadsheet, graphics, modelling) is being developed far more quickly on mainframes. Our emphasis should be on interface techniques.
- i. There is money to be made by selling data interactively to the private sector. This money will help to cover the cost of data storage and computing and contribute to future development costs. This is particularly important when government is reducing the support given to academic and research institutions.
- j. Software must be portable across computers. Unix has established itself as market leader and the language C may be even more important. Software should not be constrained to operate within the current limitation of memory/disk of today's micros.
- k. In view of the data compression techniques

now available, it is possible to store and analyse very large databases on micros and distribute the data on floppy disks. It should not be assumed that these large surveys can only be handled on large mini or mainframe computers.

- l. More research needs to be put into Expert systems that ask users what they want. The software then does the searching and decision making jointly with the user.

Ignoring change will not make it go away. Quanline has shown that the trend is to interactive tabulations, a reduction in printed reports resulting in greater freedom and wider distribution of survey data. Archivists are in a unique position to beat the rest of the world.

We believe that in the long term, the concept of transposed files will become part of conventional DBMS technology, providing the benefits of data compression and fast access for ad hoc interrogations while preserving fast retrieval and update.

The concept of interactive access to survey data will become a small but vital part of the technique of converting survey data into useful information.

Background Information

QUANTIME is a major software and systems house serving the market research industry, with over 70 people worldwide, 50 user sites and some hundreds of clients using Quantum/Quanvert to analyse survey data. Quantime's headquarters are in central London, with offices in New York and Cincinnati and major agencies in Europe. Most of Quantime's work is for the private sector, but increasingly

the public sector is taking advantage of these services.

Almost all of Quantime's development and services are based on DEC/VAX's running under Unix – in fact there are six VAX/750's spread around the world linked with a sophisticated network of telecommunications hardware and software. Quantime is both a developer and user of software, offering a tabulation bureau service, time-sharing and the sale of software and hardware. Software includes highly specialised tools for Computer Assisted Telephone Interviewing, Automatic Questionnaire printing and direct data entry – all closely integrated with data editing and analysis packages.

In 1984, Quantime opened a new division QUANLINE dedicated to the needs of users wishing to load and access remote survey databases. This is based on two of the computers and currently hosts some forty databases, requiring 1,600 Mbytes of disk storage.

Quanvert is available in two ways:

- As a software package in its own right for use on IBM mainframes and PC's, DEC/VAX, PRIME and many other Unix-based minis. Normally, one would take QUANTUM (and FLIP) in order to be able to set up the Quanvert databases. However, where a user wishes to distribute databases, Quantime also supplies a 'read only' version of Quanvert.
- Through the Quanline time-sharing service. The UK operation operates from London and a new US service will be launched in the summer from Quantime's New York office.

Appendix

The Transposed File Concept

- What is it?

A conventional data file may be considered as a matrix, with records as the rows and variables as columns/fields. Any analysis of this file involves scanning sequentially through the matrix but this is wasteful since it is unlikely that any tabulation needs to read ALL records and ALL variables.

There are a number of techniques to minimise the time to isolate and select pre-specified records – these include index sequential or random access, heaps, lists and overflows, but they all demand a choice by the user of key variables – a choice that may be difficult to make.

The concept of a transposed file is the conversion of the data file into a set of smaller files, one for each variable. These files are (unlike a pure relational database) simply sequential files holding the response from each case as a single record. It is not linked in any way to other files – the relationship between them is purely positional, i.e. the 412th record occupies the same logical position in each file.

To prepare the transposed file, a special program is run, which reads through the data file sequentially and write out a series of subfiles. This is a once only process which requires both the data file and a copy of the variable description file. When this transposed file has been created, there is no further use for the original data and description files.

- The benefits

Any program wishing to read the data, need only pull off the relevant variable files, normally a very small subset of the full data file.

The benefit of this approach is that any analysis of the data file can be very fast. It is a function of the number of records, NOT the size of datafiles. Furthermore, all variables have equal importance – there is no need for the user to nominate key variables when setting up the database. But there is another major benefit – data compression. Since each subfile contains values of only one variable, there is significant scope for data compression.

- Data compression

There are four main opportunities for such compression:

- Where there are frequently repeated values, e.g. if the data is grouped in geographic order and the first 1,000 records relate to people in Scotland, the next 3,000 in Wales, etc.
- Where data are 'missing', e.g. the variable *salary* only has values for employed people.
- Where only a few records have a specific value, e.g. if on the file only 10% of all people are women.
- Where the data are hierarchical, there is no need to repeat variables at a higher level for variables at a lower level.

There are more advanced methods for data compression using HUFFMAN coding and pattern searching, which is fairly easy to achieve on such files. The use of MONTE

CARLO simulations has revealed some useful statistics about the repeatability of bit strings on these types of transposed files.

The result of this is that data storage requirements can be reduced dramatically and data reading time reduced accordingly. Compression ratios of over 50% are often achieved and it is not unusual to see figures in excess of 95% for specific variables. The size of the file, in many cases is less than the size of the raw questionnaire-based data and further research is being carried out to improve these ratios. ▢

News & Notes

1985 Annual Program Report ¹

Over 75 people attended the Section's Program at the 1985 ALA Annual Meeting in Chicago. The panel, held at the Palmer House on Sunday, July 7, was on "Machine-Readable Data Files for Social Science: The Librarian's Role." Four panelists each addressed four questions: the value of numeric databases for social science in general and anthropology and sociology in particular, types of users (real or potential), whether such databases and support staff should be housed in libraries, and what the implications of changing computer technologies are for database access by end users — would libraries and librarians continue to play a role?

Two panel members represented social science database vendors: Larry Carbaugh of the Data User Services Division of the Bureau of Census, and Carolyn Geda of the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan. The other two panelists were librarians from universities where libraries play a central role in providing access to those and other numeric databases: Bliss Siman from Baruch College of the City University of New York and Barbara Wittkopf of the University of Florida. The panel was organized by the ANSS Chicago Program Planning Committee: Virginia Moreland of the University of Nebraska at Lincoln, Fred Peal of New York University, and Co-Chairs Janet Steins of SUNY, Stony Brook and Gregory Finnegan of Roosevelt University. Finnegan moderated the panel's discussion.

The panelists devoted most of their time to a discussion of the range of services provided by their organizations and the contributions that librarians make to those services. ICPSR communicates to members of the Consortium through a network of official representatives from each member institution. While recognizing that libraries and librarians should have a central and visible role in the provision of access to data by scholars, it is more typical that the various official representatives are members of the faculty of an academic department. The University of Florida's service has evolved from handling machine-readable Census data in the Reference Department to becoming a central access point to a great variety of data from on-campus and off-campus sources, including ICPSR.

¹Reprinted from the ANNS Currents, The newsletter of the ACRL Anthropology & Sociology Section.

The issues of who should use MRDF and whether libraries should provide access were considered secondary by the panel. In a larger context, no one disputed the advantages libraries have in providing access to all media of information. Vendors and librarians alike also had a strong sense of the utilization of MRDF at all levels of teaching as well as research.

Major emphasis was also placed on the impact of new technologies. Librarians have just begun to assimilate MRDF into the mainstream of their collections and services, and yet new means of distribution of data such as diskettes and optical disks promise to put data directly into the hands of end-users in the near future. While panelists were sensitive to this development, they generally felt that in the realm of large data sets the limited storage capacity of diskettes meant that publication per se in that medium was not a great step forward. An example of this limitation is the recently released diskette version of the "City and County Databook", which occupies 33 floppy disks. The existence of a variety of microcomputer operating systems compounds this problem. What the panelists did see as a great step forward was the potential to customized datasets for individual users by downloading from large, tape-based sets. This in turn means that librarians remain as brokers between masses of data (in all media) and the patron's specific needs. On the last point, Barbara Wittkopf responded to a question about librarians' statistical competencies by noting that machine-readable research retains the equivalent of "reading the book" — we provide access to information, not finished reports.

The program was tape-recorded, and is available from ALA.

Gregory Finnegan
Dartmouth College

Contents of Current Journals

Humanistiske data 3 86

(Norwegian Computing Centre for the Humanities)

Articles

- p. 4 Computer communication./ E.S. Ore [in Norwegian]
- p. 13 Humanities data bases./ K. Natvig [in Norwegian]
- p. 22 Computing and social studies./ H. Johansen [in Norwegian]
- p. 27 Machine-readable texts in English language research./ Stig Johansson
- p. 35 Computers and philosophical manuscripts./ C. Huitfeldt

Reports

- p. 46 nota bene/ T. Seland [in Norwegian]
- p. 53 The Centre's videodisc project – a status report./ K. Natvig [in Norwegian]
- p. 56 The Computer Secretariate and the humanities./ K. Natvig [in Norwegian]
- p. 57 Norwegian data base of place names./ B. Helleland [in Norwegian]
- p. 60 IBM seminar on speech processing./ K. Slethei [in Norwegian]
- p. 62 Course in the design of educational software./ K. Natvig [in Norwegian]
- p. 65 The Computing Secretariate's seminar in Halden./ E.S. Ore [in Norwegian]
- p. 66 The humanities and knowledge-based systems./ E. Fjornes [in Norwegian]
- p. 67 Communication and computing technology as an infrastructure in northern Norwegian
business and administration./ F. Fagertun [in Norwegian]
- p. 72 Nordic seminar on machine translation./ J.H. Hauge [in Norwegian]
- p. 81 News from the Council for Research in the Humanities. [in Norwegian]
- p. 97 News
- p.107 Summary [in English]

DDA nyt nr. 37 foraar 1986
(Dansk Data Arkiv/Danish Data Archives)

- p. 3 Editorial [in Danish]
- p. 5 Tabulations on the DDA study descriptions./K.B. Rasmussen
- p. 24 New catalogues from DDA's foreign 'sibling'-archives. [in Danish]
- p. 26 Secondary analysis of survey data. [in Danish]
Recently processed data files [in Danish]
- p. 33 DDA-0061 Social and environmental conditions in rural districts, 1973-1976.
DDA-0772 Danish election study, 1984.
DDA-0812 Industrial closedowns: Kastrup-Holmegaard 1978-1980.
DDA-0871 Attitudes and opinions of young people, March-April 1984
DDA-1192 Referendum on the EEC package, 27 February 1986.
Recently located data files [not held by DDA] [in Danish]
DDA-0755 Koldinghus fief ledgers, 1610-1611.
DDA-0846 Price material from Hamburg, 1850-1863.
DDA-0865 Noise-affected dwellings around Kastrup airport, 1983.
DDA-0866 Dietary habits and health conditions.
DDA-0867 Function of the appeal system in social service cases II, 1980.
DDA-0868 New production techniques and in-service training.
DDA-0869 Demand for manpower and staff recruitment, 1985.
DDA-1065 Working conditions and pregnancy, 1980.
- p. 48 English summary

Survey methodology, a journal of Statistics Canada
vol.12(1) June 1986

Special issue: missing data in surveys.

- p. 1 The treatment of missing survey data./ G. Kalton and D. Kasprzyk
- p. 17 On the definitions of response rates./ R. Platek and G.B. Gray
- p. 29 Some optimality results in the presence of nonresponse./ V.P. Godambe and M.E. Thompson
- p. 37 Basic ideas of multiple imputation for nonresponse./ D.B. Rubin
- p. 49 Imputation options in a generalized edit and imputation system./ P. Giles and C. Patrick
- p. 61 The maximum likelihood method for nonresponse in sample surveys./ M.S. Srivastava and E.M. Carter
- p. 73 Statistical editing and imputation for periodic business surveys./ M.A. Hidioglou and J.-M. Berthelot
- p. 85 Practical criteria for definition of weighting classes./ V. Tremblay
- p. 99 A study of the effects of imputation groups in the nearest neighbour imputation method for the national farm survey./ S. Cheung and C. Seko

Machine Readable Archives. Bulletin vol.4(3), Fall 1986
(Machine Readable Archives. Public Archives of Canada)

- p. 1 Making census manuscript schedules machine-readable: industry in Ontario urban centres, 1870./E. Bloomfield

European political data newsletter no.61, December 1986

Data section:

- p. 5 Social science data services: new data catalogues on data holdings

Books:

- p. 17 Guide to British government.
p. 18 The Asian political dictionary.
p. 19 The Gallup survey of Britain.
p. 21 How Britain votes.
p. 22 The state of Britain.
p. 23 Reference books on local elections in England.
p. 24 History and computing.
p. 26 The state in global perspective.
p. 27 Systems of geographical information.
p. 28 Space data matrices and computer graphics.
p. 30 European parliamentary election results.
p. 32 Economic sanctions reconsidered.

Computer section:

- p. 34 ADMS: aggregate data management in statistics and planning/ Oswald Schechtner and Karl Zelle
p. 41 Microcomputer software packages produced by organizations of the United Nations system.
p. 47 Programming SAS for causal analysis.
p. 49 Statistix: an interactive statistical program for microcomputers.

Forthcoming events:

- p. 52 IASSIST 1987 - 13th annual conference of data archive and data library professionals.
p. 54 ICDBHSS 1987 - international conference on data bases for the humanities and social sciences.

Historical methods
vol.20(1), Winter 1987

- p. 5 The press as a source of socio-historical data: issues in the methodology of data collection from newspapers./ Roberto Franzosi
- p. 17 The life expectancies of colonial Maryland legislators./ Daniel S. Levy
- p. 29 Analysis of reconstituted families: a package of SAS programs./ Rene Leboutte, George Alter and Myron Gutmann
- p. 35 The handwriting on the screen./ Richard Jensen

ESRC data archive bulletin nr. 36, January 1987

News:

- p. 1 Acknowledging use of data.
British social attitudes survey: a new service.
Bulletin board.
Workshop on administrative data.
Iron out your information retrieval problems with IRON.
- p. 2 Domesday disks launched.
- p. 3 Meeting on longitudinal studies.
The GHS Newsletter.
Computer liaison officers.
- p. 4 An anniversary and a contest.
Large-scale data resources for the social sciences.
- p. 5 Rural areas database.
Bulletin back issues.
New acquisitions:
- p. 5 Notes on appendices.

Selected new acquisitions:

- European communities studies, 1973-1984 (ICPSR 8434)
- p. 6 Port Talbot people and jobs survey, 1984.
(C.C. Harris, University college of Swansea)
Blackburn and Darwen labour market survey data, 1985. (P.J. Halfpenny, University of Manchester)
- p. 7 Updates to serial holdings.

Research organisations, data institutions and foreign archives:

- p. 7 ICPSR news.
Spanish data bank.
Public archives Canada.
- p. 8 Steinmetz archives catalogue
New DDA catalogue.
"NCDS V: the cohort and their children", by John Fox
Software bulletin

Notes:

- p. 10 Statistical news.
Europe as seen by Europeans.
- p. 11 Employment gazette.
European political data.
ESRC news.
AIDS database.
Medical sociology news.
Market research society "Then and now".
- p. 12 Women in the labour force.
Survey methods newsletter.
International social science journal.
Essex papers in politics and government.
- p. 13 Forthcoming events.

Books:

- p. 19 British social attitudes: the 1986 report, edited by Roger Jowell, Sharon Witherspoon and Lindsay Brook
An introduction to rural geography, by A. Gilg (reviewed by N. Walford)
- p. 20 Contrasting values in western Europe, by Stephen Harding and David Phillips, with Michael Fogarty
Key data, from the Central statistical office.
British workplace industrial relations 1980-84, by Neil Millward and Mark Stevens.
Social science and social policy, edited by Martin Bulmer.
- p. 21 Political communication: the general election campaign of 1983, edited by Ivor Crewe and Martin Harrop.
Television coverage of the 1983 general election, by Barrie Gunter, Michael Svennevig and Mallory Wober.
Images of social stratification, by A.P.M. Coxon and P.M. Davies.
New tools for social scientists, edited by William D. Berry and Michael S. Lewis-Beck.
Validity in survey research, by William A. Belson.
Proceedings of the Fifth British National Conference on Databases (BNCODS), edited by E.A. Oxborrow.
- p. 22 Directory for PC users.

Appendices:

- p. 22 A: recent additions to the ICPSR archive at Michigan
- p. 23 B: Datasets acquired since the publication of bulletin 35

Survey Methodology: a journal of Statistics Canada vol.12(2) Dec. 1986

- p.107 Estimating a monthly index based on trimestrial data./ J.G. Kovar
- p.121 Regression analysis using survey data with endogenous design./ A. Ten Cate
- p.139 A cluster analysis of activities of daily living from the Canadian Health and Disability Survey./ D.A. Binder and G. Lazarus
- p.151 Additive versus multiplicative seasonal adjustment when there are fast changes in the trend-cycle./ G. Huot and N. Gait

Special section - missing data in surveys:

- p.161 Nonresponse adjustment procedures at the U.S. Bureau of the Census./ D.W. Chapman, L. Bailey, and D. Kasprzyk
- p.181 Hot deck imputation procedure applied to a double sampling design./ S. Hinkins and F. Scheuren
- p.197 Comparison of weighting and imputation methods for estimating unsampled data./ S. Michaud
- p.207 A regression approach to estimation in the presence of nonresponse./ C.E. Særndal
- p.217 Ratio estimation with subsampling the nonrespondents./ P.S.R.S. Rao
- p.231 Acknowledgements

Computers and the social sciences vol.2(4) Oct-Dec 1986

- p.183 Computing and the political world./ James N. Danziger
- p.201 Technological determinism in social data analysis./ Martin L. Levin
- p.209 Acceptance of computer-based models in local government: information adequacy and implementation./ Susan H. Komsky
- p.221 *Book reviews:*
Ulrich Briefs, John Kjaer, and Jean-Louis Rigal, eds.,
Computerization and work: a reader on social aspects of computerization (Paul Attewell)
Marvin B. Sussman, ed., Personal computers and the family (Alladi Venkatesh)
Joan Frye Williams, ed., Online catalog screen displays: a series of discussions (Richard Ziegfield)
- p.227 *Software reviews:*
Nota Bene: powerful text processing for academics (Rodney Muth)
Energraphics, version 1.3 (Ruth S. Brent)
Chart-Master, version 6.1 (Ruth S. Brent)

Diagram-Master, version 5.0 (Ruth S. Brent)
 Map-Master (Ruth S. Brent)
 The integrated bibliographic software system:
 Pro-Search, Biblio-Link and Pro-Cite (James D. Campbell)

Historical social research

vol. 39, July 1986

- p. 3 Continuity and change in the recruitment of SPD members in a Berlin region, 1945 to 1973/
J.-B. Hohmann, H. Hurwitz & G. Kuckhahn [in German]
- p. 36 Social situation and political orientation - students and professors at Giessen University
1918-1945. Part two./ P. Chroust
- p. 86 Breaking of social barriers as an expression of the emergence of a modern society in the
mid-19th century - based on the example of selected Polish towns/ W. Molik, K.
Makowski
- p.101 The defeat of the German universities 1933/ B. W. Reimann
- p.106 *Book reviews:*
 Berg, Werner/ Wirtschaft und Gesellschaft in Deutschland und Grossbritannien im
 Uebergang zum 'organisierten Kapitalismus'. Angestellte, Arbeiter und Staat des
 Ruhrgebietes und von Suedwales, 1850-1914. 1984. [in German]
 Cronin, James E./ Labour and society in Britain 1918-1979. 1984
 Hinton, James/ Labour and socialism. A history of the British labour movement. 1983
 Pimlott, Ben & Chris Cook (eds.)/ Trade unions in British politics. 1982
- p.113 Data news [in French]
- p.117 Quantum information
- p.124 Forthcoming conferences
- p.126 Publication notices

social research

vol. 40, October 1986

- p. 3 A draft proposal for a standard for the coding of machine readable sources. / Manfred
Thaller
- p. 47 The development of the politics of housing up to WW1. / Elisabeth Gransche [in German]
- p. 72 Legal transcripts of the Bremen lawcourts, 1600-1800 / Peter Kottmann [in German]
- p. 84 *Book review:*
 Nicosia, Francis R./ The Third Reich and the Palestine question. 1985
- p. 88 *Software:*
 SPSS/PC: a quantitative historian's dream or nightmare?/ Konrad H. Jarausch
- p. 91 Once more into the breach: computer literacy and the humanities / Kevin Roddy

Data news:

- p. 96 Exaggerated data protection hinders historical research/ Jurgen Kocka [in German]
- p.103 Quantum information
- p.107 Forthcoming conferences
- p.113 Position available [in German]
- p.118 Cumulative contents HSR vol. 37 – 40

Machine Readable Archives. Bulletin vol.4(4), Winter 1987 (Machine Readable Archives. Public Archives of Canada)

- p. 1 Organizational changes.
The appraisal of government EDP records/ David Brown & Katharine Gavrel



I A S S I S T

The International Association for Social Science Information Services and Technology (IASSIST) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the IASSIST QUARTERLY. They also benefit

from reduced fees for attendance at regional and international conferences sponsored by IASSIST.

Membership fees are:

Regular Membership: \$20 per calendar year
Student Membership: \$10 per calendar year

International subscriptions to the QUARTERLY are available, but do not confer voting rights or other membership benefits.

Institutional Subscription: \$35 per calendar year
(includes one volume of the QUARTERLY)

I would like to become a member of IASSIST. Please see my choice below:

- ☐ \$20 Regular Membership
☐ \$10 Student Membership
☐ \$35 Institutional Membership

My primary interests are:

- ☐ Archive Services/Administration
☐ Data Processing/Data Management
☐ Research Applications
☐ Other (specify) _____


Name Phone


Institutional Affiliation

Mailing Address

City

Country zip/postal code

 Please make checks payable to IASSIST and Mail to :

 Ms. Jackie McGee
Treasurer, IASSIST
Rand Corporation
1700 Main Street
Santa Monica, CA. 90406



LIBRARY

APR 8 1988

School of Library Science
University of North Carolina